# [WEDT.A] Klasyfikacja typów serwisów WWW na postawie informacji o strukturze strony i tekstu

Michał Aniserowicz, Jakub Turek

## Opis problemu

Zadanie polega na implementacji aplikacji, która dokonuje automatycznej klasyfikacji typów stron WWW na podstawie ich struktury. Analiza może obejmować źródło strony, konfigurację rozmieszczenia komponentów (layout), a także strukturę i znaczenie zamieszczonych na stronie treści.

## Założenia

Projekt obejmuje implementację klasyfikatora następujących typów serwisów:

**Blog** rodzaj internetowego dziennika (pamiętnika), który zawiera odrębnie, chronologicznie uporządkowane wpisy. Przykład serwisu: `http://rafalstec.blox.pl/`.

**Serwisy informacyjne** portale zawierające najnowsze wiadomości z różnych dziedzin życia, takich jak polityka, finanse, technologie. Przykład serwisu: `http://onet.pl/`.

**„Kwejki”** serwisy społecznościowe oparte w głównej mierze na grafikach. Przykład serwisu: `http://kwejk.pl/`.

**Sklepy internetowe** portale umożliwiające wyszukiwanie najtańszych przedmiotów w ramach zadanych parametrów. Przykład serwisu: `http://ceneo.pl/`.

Dane wejściowe aplikacji stanowić będzie adres witryny internetowej. Na wyjście wyprowadzona zostanie nazwa kategorii lub informacja, że serwis nie został zaklasyfikowany do żadnej z powyższych kategorii.

## Struktura danych

Aplikacja umożliwiać będzie budowanie pełnego drzewa HTML. W korzeniu drzewa przechowywane będą następujące informacje:

- Typ napotkanego taga HTML, na przykład `<div>`, `<h1>`.

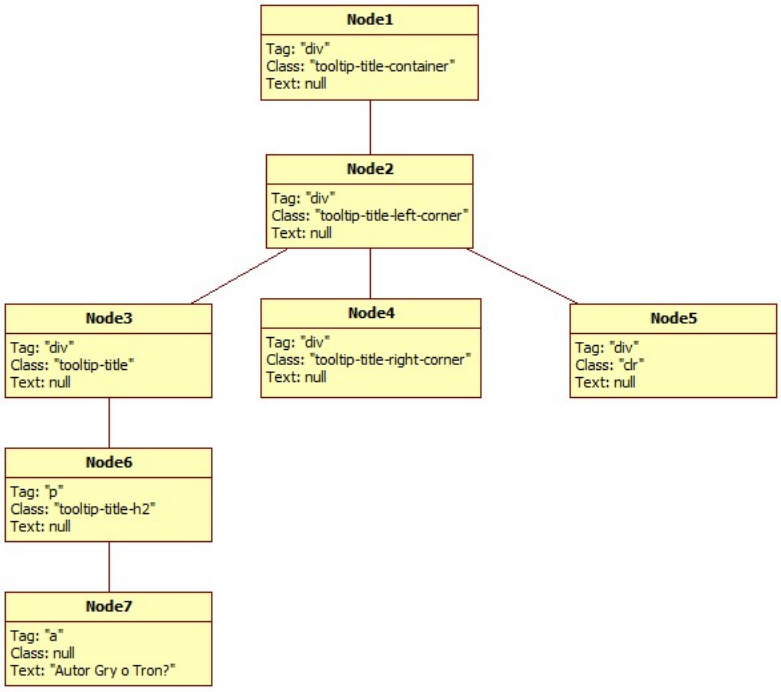
```
<div class="tooltip-title-container">
  <div class="tooltip-title-left-corner">
    <div class="tooltip-title">
      <p class="tooltip-title-h2">
        <a href="/obrazek/1763501/autor-gry-o-tron.html">
          Autor Gry o Tron?
        </a>
      </p>
    </div>
    <div class="tooltip-title-right-corner"></div>
    <div class="clr"></div>
  </div>
</div>
```

Rysunek 1: Fragment kodu źródłowego witryny http://kwejk.pl.

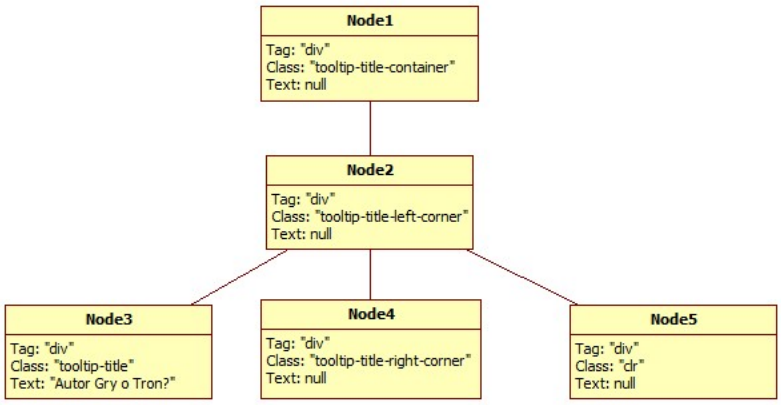
- Dodatkowe atrybuty taga powiązane z CSS - kaskadowymi arkuszami styli: identyfikator `id="_"`, klasa `class="_"` oraz styl elementu `style="_"`.
- Tekst zawarty pomiędzy tagiem otwierającym a zamykającym. Przykładowo dla kodu `<a>Odnośnik</a>` jest to fraza „Odnośnik”.
- Inne atrybuty kontekstowe związane z poszczególnymi tagami:
  - dla obrazka (`<img>`) - jego rozmiar oraz źródło pochodzenia (lokalne - z domeny, którą analizujemy lub zewnętrzne - spoza niej),
  - dla nagłówków (`<h1>`, `<h2>`, itd.) - rozmiar czcionki.

Ze względu na rozmiary oraz skomplikowanie struktury dla dużych portali, takich jak sklepy internetowe lub serwisy informacyjne, kod aplikacji będzie udostępniał różne możliwości redukcji złożoności drzewa:

- Zawężanie podzbioru tagów, dla których budowane jest drzewo. Tagi istotne dla struktury strony to, między innymi, `<div>`, `<td>`, `<article>`, `<h1>`, `<a>` oraz `<img>`. Z punktu widzenia zadania, tagi niosące niewiele informacji służą głównie do formatowania tekstu, jak na przykład `<b>`, `<span>`, oraz osadzania skryptów - `<script>`.
- Ograniczanie stopnia zagnieżdżania korzeni w drzewie:
  - pomijanie węzłów przekraczających dany, parametryzowalny, poziom zagłębienia w strukturze,
  - sklejanie kilku następujących po sobie węzłów o zbliżonych wymiarach na stronie w jeden.
- Odfiltrowywanie elementów uznanych za nieistotne metodami heurystycznymi, przykładowo prosty filtr eliminujący reklamy bazując na klasach obiektów.




Rysunek 2: Pełne drzewo HTML dla kodu przedstawionego na listingu 1.



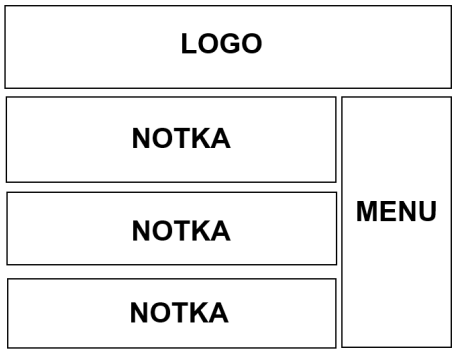

Rysunek 3: Drzewo HTML z rysunku 2 zredukowane do tagów div.

# Algorytm

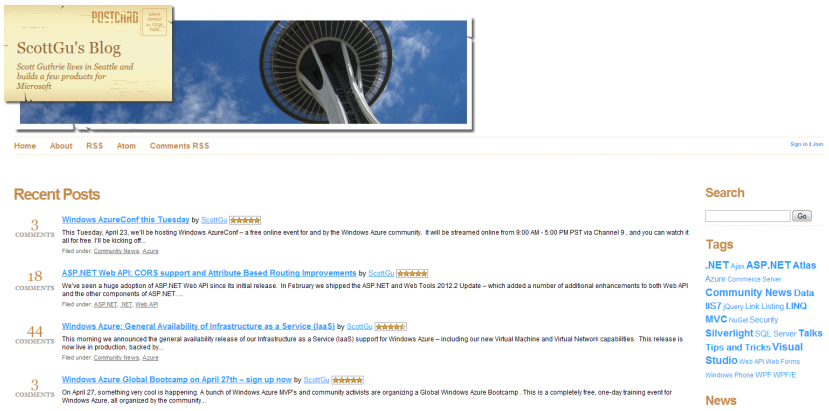
## Blogi

Algorytm kategoryzacji blogów będzie rozpoznawał dzienniki internetowe w dwukolumnowym układzie. W szerszej kolumnie znajdują się notki, natomiast węż-

sza kolumna to menu strony. Notki posiadają pewną stałą strukturę wewnętrzną. Na pojedynczy wpis składają się: tytuł, treść, data, informacja o autorze oraz link do komentarzy. Menu charakteryzuje się natomiast występowaniem dużej ilości podobnych odnośników jeden pod drugim (archiwum dla kolejnych miesięcy, kategorie wpisów, odnośniki do zaprzyjaźnionych blogów). Szablon bloga przedstawia rysunek 4. Przykłady stron o takiej strukturze przedstawiono na rysunkach 5 oraz 6.




Rysunek 4: Szablon, na którym oparta jest większość współczesnych blogów.




Rysunek 5: Przykład dwukolumnowej struktury bloga.

Algorytm rozpoznawania blogów będzie działał według następującego schematu:

1. Odnalezienie w drzewie powtarzającego się układu węzłów, które reprezentują notkę. Cechy charakterystyczne takiego układu to:
   - Duża ilość tekstu zawartego pomiędzy tagami strukturalnymi, z niewielką ilością odnośników.
   - Można wyróżnić przynajmniej tytuł notki oraz odnośnik do komentarzy.
   - Odpowiadające tagi strukturalne mają identyczne właściwości - klasy kaskadowych arkuszy styli.


Rysunek 6: Przykład dwukolumnowej struktury bloga.

- Kolejne struktury umieszczone są w kodzie jedna pod drugą oraz wszystkie mają wspólnego rodzica.
- Powtarzalność struktury wynosi co najmniej pięć (założenie heurystyczne).

2. Odnalezienie w drzewie układu węzłów, który repreprezentuje menu strony (schematyczne odnośniki występujące jeden pod drugim i prowadzące głównie do adresów w domenie lokalnej - archiwum oraz kategorie).

3. Badanie wzajemnych pozycji odnalezionych kolumn - powinny być umieszczone horyzontalnie względem siebie.

4. Odnalezienie nagłówka strony, czyli największego elementu graficznego / tekstowego na stronie, występującego pojedynczo.

5. Badanie wzajemnych pozycji przypuszczalnego nagłówka strony oraz kolumn zawierających menu i wpisy. Nagłówek strony musi znajdować się ponad pozostałymi kolumnami.

Jeżeli wszystkie elementy udało się odnaleźć oraz spełnione są przytoczone założenia przestrzenne, to witryna kategoryzowana jest jako blog.

## Strony informacyjne

Strony informacyjne charakteryzują się liczną zawartością odnośników w domenie lokalnej. Podobnie jak w przypadku blogów, dominuje układ dwukolumnowy, przy czym kolumna szersza jest zazwyczaj nieustrukturalizowana, natomiast kolumna węższa zawiera wiele odnośników jeden pod drugim. Logotyp

nie jest dominującym elementem strony i ciężko wyznaczyć go na bazie wielkości, zazwyczaj położony jest w lewym górnym rogu witryny. Ponadto, strony informacyjne charakteryzują się wykorzystaniem trzypoziomowej strukturyzacji oferowanej przez HTML5, za pomocą tagów `<div>`, `<article>` oraz `<section>`.

Algorytm rozpoznawania stron informacyjnych będzie działał według następującego schematu:

1. Zbadanie proporcji zwykłego tekstu występującego w ramach odnośników (`<a>...</a>`) oraz innych tagów zawierających tekst widoczny na witrynie.

2. Określenie częstotliwości wykorzystania tagów `<article>` oraz `<span>` do strukturyzacji zagnieżdzonych odnośników do domen lokalnych (newsy).

3. Odnalezienie wąskiej kolumny zawierającej odnośniki do wiadomości.

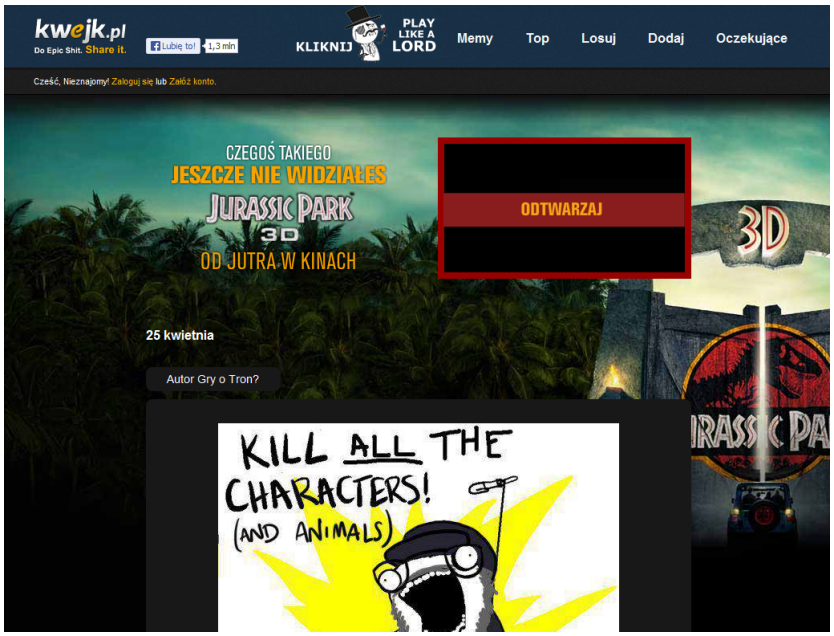
Jeżeli proporcje w punkcie 1. są $\gg 1$ oraz częstotliwość w punkcie 2. jest wysoka, a także udało odnaleźć się kolumnę z punktu 3. to strona klasyfikowana jest jako strona informacyjna.

## „Kwejki"

Strony typu *kwejk* charakteryzują się występowaniem pojedynczej, centralnej kolumny, która zawiera dużych rozmiarów obrazy umieszczone jeden pod drugim. Obrazy te są najczęściej prezentowane w ramach jednego „pudełka", a więc są osadzone w regularnej strukturze tagów HTML. Ponad obrazami znajduje się logotyp oraz menu strony, natomiast pod nimi umieszczone jest małe menu nawigacyjne, prezentujące kolejne liczby (strony). Ogólny szablon stron typu *kwejk* przedstawia rysunek 7, natomiast rysunki 8 oraz 9 pokazują przykłady takich stron.


Rysunek 7: Szablon stron typu *kwejk*.

Algorytm kategoryzacji „kwejków" polega na odnalezieniu powtarzającej się struktury węzłów, o wspólnym rodzicu i identycznych właściwościach, w każdym z których osadzona jest pojedyncza grafika. Jeżeli taka struktura występuje, badane jest położenie tej struktury na stronie. W przypadku, gdy kolumna jest wyśrodkowana, strona może zostać zakwalifikowana jako „kwejk".




Rysunek 8: Portal społecznościowy http://kwejk.pl oparty na grafikach.

## Sklepy internetowe


Rysunek 9: Portal społecznościowy http://demotywatory.pl oparty na grafikach.